Edition HMD

Daniel Fasel
Andreas Meier *Hrsg.*

Big Data

Grundlagen, Systeme und Nutzungspotenziale

Praxis der Wirtschaftsinformatik

HMD

Springer Vieweg

Edition HMD

Die Reihe Edition HMD wird herausgegeben von Hans-Peter Fröschle, Prof. Dr. Knut Hildebrand, Dr. Josephine Hofmann, Prof. Dr. Matthias Knoll, Prof. Dr. Andreas Meier, Stefan Meinhardt, Dr. Stefan Reinheimer, Prof. Dr. Susanne Robra-Bissantz und Prof. Dr. Susanne Strahringer.

Seit über 50 Jahren erscheint die Fachzeitschrift „HMD – Praxis der Wirtschaftsinformatik“ mit Schwerpunktausgaben zu aktuellen Themen. Erhältlich sind diese Publikationen im elektronischen Einzelbezug über SpringerLink und Springer Professional sowie in gedruckter Form im Abonnement. Die Reihe „Edition HMD“ greift ausgewählte Themen auf, bündelt passende Fachbeiträge aus den HMD-Schwerpunktausgaben und macht sie allen interessierten Lesern über online- und offline-Vertriebskanäle zugänglich. Jede Ausgabe eröffnet mit einem Geleitwort der Herausgeber, die eine Orientierung im Themenfeld geben und den Bogen über alle Beiträge spannen. Die ausgewählten Beiträge aus den HMD-Schwerpunktausgaben werden nach thematischen Gesichtspunkten neu zusammengestellt. Sie werden von den Autoren im Vorfeld überarbeitet, aktualisiert und bei Bedarf inhaltlich ergänzt, um den Anforderungen der rasanten fachlichen und technischen Entwicklung der Branche Rechnung zu tragen.

Weitere Bände in dieser Reihe: http://www.springer.com/series/13850

Daniel Fasel • Andreas Meier
Herausgeber

Big Data

Grundlagen, Systeme und Nutzungspotenziale

Springer Vieweg

Herausgeber
Daniel Fasel
Scigility AG
Zürich, Schweiz

Andreas Meier
Institut für Informatik
Univ. Fribourg
Fribourg, Schweiz

Das Herausgeberwerk basiert auf Beiträgen der Zeitschrift HMD – Praxis der Wirtschaftsinformatik, die entweder unverändert übernommen oder durch die Beitragsautoren überarbeitet wurden.

ISSN 2366-1127 ISSN 2366-1135 (electronic)
Edition HMD
ISBN 978-3-658-11588-3 ISBN 978-3-658-11589-0 (eBook)
DOI 10.1007/978-3-658-11589-0

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Vieweg

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Vieweg ist Teil von Springer Nature
Die eingetragene Gesellschaft ist Springer Fachmedien Wiesbaden GmbH

Vorwort

Big Data ist zum Hype geworden. Täglich werden in den Medien Erfolgsmeldungen veröffentlicht. Blogger streiten über die Vor- und Nachteile des Einsatzes von NoSQL-Datenbanken & Co., Führungsgremien stehen unter Druck, ihre Informatikbudgets nach oben anzupassen und in Big-Data-Technologien zu investieren. Politiker fordern regionale oder nationale Programme, auf den Big-Data-Schnellzug aufzuspringen und den Einsatz für Verkehrsregelung, Energieverteilung, Wasserversorgung etc. zu prüfen. In Universitäten und Fachhochschulen schließlich wird debattiert, spezifische Studiengänge für Data Science aufzuziehen.

Was ist Big Data? Damit werden Datenbestände bezeichnet, die aufgrund ihres Umfangs (Volume), ihrer Strukturvielfalt (Variety) und ihrer Volatilität und Verfügbarkeit (Velocity) nicht in herkömmlichen, sprich relationalen Datenbanken, gehalten und eventuell nicht mit SQL (Structured Query Language) ausgewertet werden können. Sobald Firmen oder Verwaltungen umfangreiche Datenströme, soziale Medien, E-Mails, heterogene Dokumentensammlungen etc. gezielt auswerten wollen, müssen sie auf NoSQL-Technologien zurückgreifen. NoSQL steht hier für ‚Not only SQL'.

Das folgende Herausgeberwerk entstand aufgrund der HMD – Praxis der Wirtschaftsinformatik und der Publikation des Schwerpunktheftes Big Data (Band 51, Heft 4, August 2014, Springer Vieweg). Einige wenige Inhalte konnten übernommen und erweitert werden. Ein Großteil der Kapitel wurde hingegen neu konzipiert und von namhaften Experten aus Hochschule und Praxis mit aktuellem Inhalt gefüllt.

Das Werk ist in drei Teile gegliedert: Im ersten Teil I – Grundlagen – werden die Begriffe Big Data und NoSQL erläutert und die Technologien auf Reife und Nützlichkeit hin eingeschätzt. Sie entdecken hier u. a., welche Unterschiede zwischen ACID (Atomicity, Consistency, Isolation, Durability) und BASE (Basically Available, Soft State, Eventually Consistent) zur Konsistenzgewährung bestehen. Eine Marktanalyse bestätigt die Erwartungshaltung bezüglich der Nutzungspotenziale von Big Data. Sie finden zudem Anforderungsprofile und Einsatzgebiete für den Data Scientist. Ein spezifisches Kapitel widmet sich der Frage, wie die Privatsphäre im Zeitalter von Big Data geschützt resp. ein Eigentumsrecht an Daten durchgesetzt werden könnte. Der Exkurs in die datenschutzrechtlichen Themen zeigt, dass nach wie vor viele Fragen zu Big Data aus juristischer Sicht offen sind.

Der zweite Teil II – Systeme – gibt Ihnen einen Überblick über die wichtigsten NoSQL-Technologien und -Datenbanken. Unter anderem wird anschaulich erläutert, wie das Map/Reduce-Verfahren funktioniert und welcher Stellenwert ihm für paralleles Arbeiten zukommt. Data Warehousing sollte und kann mit NoSQL-Ansätzen erweitert werden, wobei Unterschiede zwischen Business Intelligence und Big Data auszumachen sind. Das massiv parallele Datenbanksystem Impala erlaubt Ihnen, weiterhin mit SQL analytische Arbeiten auf umfangreichen Datensammlungen leistungsstark durchzuführen. Um Kosten in der Cloud einzusparen, sollten die einzelnen Datenbankfunktionen beliebig konfiguriert und modular zusammengesetzt werden können. Falls Sie eine Big-Data-Anwendungsplattform anstreben, drängt sich eventuell SAP HANA auf.

NoSQL-Technologien alleine bringen Sie in Ihrem Berufsalltag nicht weiter. Aus diesem Grunde haben wir den Teil III – Nutzung – konzipiert und sind glücklich, einige interessante Anwendungsfälle aus Praxis und Wissenschaft vorstellen zu können. Als Einstieg zeigt das Kapitel über Cloud Service Management auf, wie Big Data die ITIL-Landschaft erweitert. Eine Reise in die Modellregion Salzburg lohnt sich allemal, denn hier erfahren Sie durch konkrete Anwendungsoptionen die Nutzung von Big Data in der Mobilität. Falls Sie mit den jetzigen Suchmaschinen resp. den entsprechenden Resultaten unzufrieden sind, verwenden Sie eventuell semantische Suchverfahren. Im Bereich Smart City finden Sie vielfältige Anwendungsoptionen für Big Data; ein Beispiel zur Optimierung der Wasserversorgung in Dublin soll hier das Potenzial aufzeigen. Dass die Einführung von Big Data das Unternehmen resp. die Organisation auf vielen Ebenen tangiert, illustrieren die beiden Fallbeispiele zur Migros als Detailhandelsunternehmen und zum Krankenversicherer sanitas ag. Ein weiterer Forschungsbeitrag zur Nutzung eines Granular Knowledge Cube rundet die Reihe der Fallbeispiele ab.

Die Welle von Big Data ist von den USA nach Europa übergeschwappt und die positive Einschätzung der NoSQL-Technologien wird in diesem Fachbuch bestätigt. Allerdings sind bahnbrechende Erfahrungen oder wichtige Erfolge in Unternehmen, die sich rechnen lassen, noch spärlich. Es bleibt deshalb die Hoffnung, dass die vielfältigen Potenziale für Big-Data-Anwendungen in Wirtschaft und Verwaltung breiter ausgeschöpft werden.

An dieser Stelle möchten wir uns bei all den Experten aus Forschung, Entwicklung und Praxis bedanken, die uns gemäß einem vorgeschlagenen Raster zu Grundlagen, Systeme und Anwendungen spannende Kapitel aus ihrem Erfahrungsbereich beisteuerten; eine Liste der Kurzlebensläufe finden Sie beigelegt. Zudem haben uns die Mitherausgeber der Edition HMD mit Rat und Tat unterstützt. Ein besonderes Dankeschön richten wir an Hermann Engesser von Springer Vieweg sowie an Sabine Kathke und Ann-Kristin Wiegmann, die unermüdlich das Werk betreut und viele Verbesserungen im Laufe der Zeit eingebracht haben.

Nun liegt es an Ihnen, liebe Leserinnen und Leser, sich ein kritisches Urteil zur Einschätzung von Big Data und NoSQL zu erarbeiten. Falls Sie eine Wertsteigerung für Ihr Unternehmen oder Ihre Organisation anstreben, drücken wir die Daumen.

Fribourg, im Mai 2016 Daniel Fasel und Andreas Meier

Geleitwort

Die weltweite Datenmenge explodiert.[1] Laut IBM[2] sollen 90 % der heutigen Daten erst in den vergangenen zwei Jahren entstanden sein. Eine Verzehnfachung in zwei Jahren! Diese Daten sind „big", also groß. Natürlich kann argumentiert werden, dass ein Großteil der heute neu hinzukommenden Daten Multimedia-Streams sind, also Audio und Video, welche zwar viele Bytes, aber wenig Information enthalten; und teilweise sogar völlig frei sind von anwendbarem Wissen. Dennoch bedeuten „Big Data" gesellschaftlich eine noch nie da gewesene, exponentielle Zunahme der weltweiten Datenmenge. Das heisst, dass der Anteil anwendbaren Wissens im Verhältnis zu vorhandenen Daten immer kleiner und kleiner wird. Ein Zitat von John Niasbitt macht das deutlich: „We are drowning in information but starved for knowledge."[3]

Aus technischer Sicht stellen „Big Data" die Frage nach Lösungen für das Problem der Datenflut. Es ist klar, dass „Big Data" zum Hype werden, wenn die Obama-Administration im Jahr 2012 mit der „Big Data Initiative" 200 Millionen US Dollar für die Forschung in diesem Bereich zur Verfügung stellt.[4] Die Affäre um die Internetschnüffelei im Jahr 2013 lässt das Thema allerdings in einem etwas anderen Licht erscheinen: „Die kopieren das ganze Internet", schrieb der Tagesanzeiger im Juni 2013.[5] Es ist die Rede von „Zettabytes". Die Frage drängt sich auf, ob ein Zusammenhang besteht zwischen Obama's Forschungsmittel-Giesskanne und der Tatsache, dass die unheimliche Menge an Überwachungsdaten mit herkömmlichen Methoden der Computer Science gar nicht mehr auswertbar ist. Der gläserne Bürger wird zur Nadel im Heuhaufen.

---

[1] Dieses Geleitwort entspricht dem Einwurf ‚Die Geister, die wir riefen' aus der Zeitschrift HMD – Praxis der Wirtschaftsinformatik, Band 51, Heft 4, August 2014, S. 383–385.

[2] IBM (2014). What is Big Data? http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

[3] Naisbitt, J. (1988). Megatrends. Grand Central Publishing.

[4] Executive Office of the President of the United States (2012). Obama Adminstration unveils „Big Data" Initiative. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

[5] Der Tagesanzeiger (2013). Nutzen die Spione Hintertüren zu Facebook Google und Apple? http://www.tagesanzeiger.ch/ausland/amerika/Nutzen-die-Spione-Hintertueren-zu-Facebook-Google-und-Apple/story/19734103.

Wohin wird uns die Datenexplosion führen? Wie wird sie uns verändern? In den Worten des Historikers Philip Blom[6]: „Man kann Technologie nicht gebrauchen, ohne durch sie verändert zu werden; und zwar bis ins Innerste, Intimste, verändert zu werden. (…) Wir sind mehr denn je überwältigt von Information, die wir gar nicht so schnell assimilieren können. Sie prasselt auf uns ein, 24 Stunden pro Tag. Das sind technologische Gegebenheiten, die uns verändern, und unser Verständnis hechelt dem hinterher. Wir können das gar nicht so schnell verstehen. Unser Verständnis ist linear, aber die technologische Entwicklung ist eigentlich exponentiell."

Die Daten wachsen uns über den Kopf. Die Geschichte ähnelt Goethes Zauberlehrling,[7] der aus lauter Faulheit einen Besen zum Leben erweckt, damit dieser für ihn den Boden wischt. Dieser durch Zauberei automatisierte Besen bringt immer mehr und mehr Wasser in die Zauberwerkstatt, bis diese überschwemmt wird: „Die Not ist gross! Die ich rief, die Geister, werd ich nun nicht mehr los!" Auf analoge Weise haben wir Menschen Daten verarbeitende Maschinen erschaffen, damit wir nicht mehr selbst denken müssen. Diese Maschinen schwemmen nun allerdings immer mehr und mehr Daten an.

Das Datenwachstum wird Methoden und Werkzeuge notwendig machen, um mit der Datenflut umzugehen. Das kann eine Rückkoppelung anheizen, da diese Werkzeuge selbst wieder zusätzliche Daten generieren. Wenn sich keine Sättigung einstellt, wenn die Daten und die Technologien immer weiter exponentiell anwachsen: ist dann eine technologische Singularität[8] möglich, eine künstliche Superintelligenz via progressio ad infinitum, wie es Vernor Vinge, Professor für Mathematik (und, zugegebenermaßen, Science-Fiction-Autor) im Jahr 1993 beschrieben hat…?

Der Bau von Supercomputern via parallelen Rechner-Clustern ist nur eine Variante zur Bewältigung der Datenflut. Vielleicht gibt es andere, antizyklische Lösungsansätze. Ich denke da an den Begriff der Wissenstechnologie[9,10] des Psychologen und Professors für künstliche Intelligenz, Sir Nigel Shadbolt. Die Wissenstechnologie erlaubt Individuen, direkt mit Wissen zu interagieren, in einer sublimierten Form; so werden wir von der Datenflut abgeschirmt. Es geht dann nicht mehr um Daten, sondern um anwendbares Wissen.

Luzern, im August 2015 Michael Kaufmann

[6]Philip Blom (2014). Europas Aufbruch ins Ungewisse. Interview im Schweizer Fernsehen SRF1, Sternstunde Philosophie vom 18. April, Minuten 18–19. http://www.srf.ch/player/tv/sternstunde-philosophie/video/philipp-blom-europas-aufbruch-ins-ungewisse?id=f9fecea0-59cc-4d22-ae23-bc1d7f895e6a.

[7]Goethe, J. W. v. (1797). Der Zauberlehrling. http://meister.igl.uni-freiburg.de/gedichte/goe_jw07.html.

[8]Vinge, V. (1993). The Coming Technological Singularity: How to Survive in the Post-Human Era. NASA technical Report No. CP-10129, S. 11–22. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855.pdf.

[9]Milton, N., Shadbolt, N., Cottam, H., & Hammersley, M. (1999). Towards a knowledge technology for knowledge management. International Journal of Human-Computer Studies, 51(3), 615–641.

[10]Shadbolt, N. R. (2001) Knowledge Technologies. Ingenia (8), 58–61, London: The Royal Academy of Engineering.

Die Autoren

Filip-Martin Brinkmann Filip-Martin Brinkmann ist wissenschaftlicher Mitarbeiter am Departement Mathematik und Informatik der Universität Basel. In der Forschungsgruppe Datenbanken und Informationssysteme von Prof. Dr. Heiko Schuldt beschäftigt er sich mit der verteilten Datenverwaltung in der Cloud. Zuvor war er mehrere Jahre lang als Gesellschafter eines IT-Systemhauses tätig. Dort betreute er mittelständische Unternehmen speziell in Fragen der Datenverwaltung und IT-Infrastrukturplanung. Er studierte Informatik an der Fachhochschule Nordwestschweiz (FHNW) und an der Universität Basel.

Richard Brunauer Richard Brunauer ist ausgebildeter Tiefbauingenieur. 2008 bzw. 2012 hat er sein Masterstudium für Angewandte Informatik und sein Diplomstudium für Analytische Philosophie an der Paris-Lodron Universität Salzburg abgeschlossen. Im Jahr 2015 promovierte er am Fachbereich Mathematik der Universität Salzburg in Technischer Mathematik zu einem Thema des Maschinellen Lernens und der Künstlichen Intelligenz. Ab 2008 arbeitete und leitete Richard Brunauer F&E-Projekte in den Bereichen Verkehrsregelung, Verkehrsprognosen und Luftschadstoffprognosen. Zum Einsatz kamen vor allem Methoden des Maschinellen Lernens, des Data-Minings und der Signalverarbeitung. Seit 2012 ist Richard Brunauer Mitarbeiter der Salzburg Research und arbeitet dort als Data Scientist mit Fokus auf Bewegungsdaten.

Philippe Cudré-Mauroux Philippe Cudré-Mauroux promovierte 2006 in Distributed Information Systems an der EPFL Lausanne in der Schweiz. Seit dem Jahr 2010 ist Philippe Cudré-Mauroux assoziierter Professor an der Universität Fribourg, Schweiz, nachdem er rund zwei Jahre am Computer Science & Artificial Intelligence Lab des MIT in den USA forschte. Er ist Autor zahlreicher internationaler Publikationen und ein gefragter Redner.

Djellel Eddine Difallah Djellel Eddine Difallah ist wissenschaftlicher Mitarbeiter am eXascale InfoLab. Dort forscht er u. a. zu Themen wie database systems (SciDB), systems performance evaluation (OLTPBench) und crowd-sourcing. Zudem promoviert er in Informatik an der Universität Fribourg, Schweiz.

Alexander Denzler Alexander Denzler doktoriert in Wirtschaftsinformatik an der Universität Fribourg, Schweiz. Er hat einen Bachelor in Wirtschaftswissenschaften

und einen Master in Wirtschaftsinformatik von der Universität Fribourg. Seine Forschung fokussiert sich auf die Anwendung von Fuzzy Logic, Granular Computing und Inference Systems zur Strukturierung von Wissen und der Empfehlung von Wissensträgern zur Teilung von Wissen innerhalb einer Community.

Daniel Fasel Daniel Fasel, Gründer und CEO der Scigility AG, befasst sich seit 2008 mit High Performance Computing Clustern, NoSQL Technologien und Big Data im BI Bereich. Ab 2011 hat er diese Technologien auch im produktiven BI Umfeld der Swisscom Schweiz AG eingesetzt. 2013 gründete Dr. Daniel Fasel zusammen mit Prof. Dr. Philippe Cudré-Mauroux die Scigility AG. 2012 erhielt er den Doktortitel in Wirtschaft von der Universität Fribourg. Er schrieb eine These im Bereich Fuzzy Data Warehousing. Sein profundes Wissen und die praktischen Erfahrungen, wie man Big Data in innovativen Schweizer Unternehmungen integral einsetzen kann, fließen direkt in die Scigility AG. Scigility AG verbindet in der Schweiz einmalige praktische Erfahrungen und akademisches Expertenwissen im Bereich vom Big Data Technologien, Advanced Analytics, Integration von NoSQL Technologien, rechtliche Aspekte zur Datenverarbeitung und Governance.

Ilir Fetai Ilir Fetai studierte Informatik an der Universität Basel und war davor als Software Engineer und Architekt in Bankensektor tätig. Aktuell arbeitet er als wissenschaftlicher Mitarbeiter in der Forschungsgruppe Datenbanken und Informationssysteme von Prof. Schuldt. Seine Forschungsschwerpunkte liegen in der verteilten Datenverwaltung in der Cloud. Seit 2011 ist er im Nebenamt noch als Dozent für Software Entwicklung an der Fernfachhochschule Schweiz tätig.

Christian Gügi Christian Gügi arbeitet als Solution Architect bei Scigility AG und ist für die Konzeption, Architektur und Realisierung von modernen Big-Data-Analyticsplattformen in Kundenprojekten verantwortlich. Er ist in der Schweiz einer der Vorreiter der aufstrebenden Disziplin Big Data und Gründer der Swiss Big Data User Group, das größte Schweizer Netzwerk zum Thema Big Data.

Urs Hengartner Urs Hengartner ist bei der Canoo Engineering AG in Basel tätig. Er ist Spezialist für Information Retrieval und Wissensmanagement und unterrichtet u. a. an der Universität Basel auf diesem Gebiet. Nach Abschluss der regulären Schulzeit am Realgymnasium Basel im Jahre 1976, arbeitete er in der Schweizerischen National Versicherung als Analytiker-Programmierer und System-Programmierer. Nach Studien an der ETH Zürich und Universität Zürich erlangte er 1990 das Diplom in Wirtschaftsinformatik an der Rechts- und Staatswissenschaftlichen Fakultät der Universität Zürich. Er promovierte im Jahr 1996 mit der Dissertation „Entwurf eines integrierten Informations-, Verwaltungs- und Retrieval Systems für textuelle Daten“. Neben seiner Lehrtätigkeit an der Uni Basel hielt er Vorträge im Bereich des Information Retrievals, Wissensmanagement und Software Engineering.

Olivier Heuberger-Götsch Olivier Heuberger-Götsch ist Rechtsanwalt und insbesondere im ICT- und Immaterialgüterrecht tätig, mit Schwerpunkten im

Persönlichkeits- und Datenschutzrecht sowie im Lizenz-, Internet- und Softwarerecht. Als Legal Counsel der Scigility AG berät er Unternehmen, Privatpersonen und Behörden in den Bereichen ICT/IP und Vertragsrecht. Zugleich verfasst Olivier Heuberger an der Universität Luzern seine Dissertation im IT-Recht.

Marcel Kornacker Marcel Kornacker ist Chefarchitekt für Datenbank-Technologien bei Cloudera und Gründer des Cloudera Impala Projektes, das heute die wohl schnellste SQL Engine für Big Data auf Hadoop ist. Nach der Promotion im Jahr 2000 im Bereich Datenbanken an der UC Berkeley hielt er diverse Positionen im Datenbankbereich bei verschiedenen Startups. Im Jahr 2003 wechselte Marcel zu Google, wo er im Bereich der Onlinewerbung und Speicherinfrastruktur arbeitete. Später wurde er der Tech Lead für den Bereich der verteilten Abfrageausführung in Google's F1 Projekt.

Dirk Kunischewski Dirk Kunischewski ist Spezialist im Bereich Business Intelligence, Datenbanken und Data Warehouse. Er verfügt über mehr als 10 Jahre praktische Erfahrung mit der Entwicklung und dem Betrieb von DWH/BI-Systemen bzw. Java-Anwendungen, die er bei der Begleitung von Projekten in der internationalen Finanz- und Kundendienstleistungsindustrie sammeln konnte.

Sean A. McKenna Sean A. McKenna arbeitet als Senior Research Manager im Bereich Constrained Resources und Environmental Analytics bei IBM Research in Dublin, Irland. Zudem ist er Research Industry Specialist (RIS) im Bereich Smart Cities. Seinen Doktortitel erhielt Sean A. McKenna 1994 als Geoingenieur an der Colorado School of Mines, USA. McKeena ist Autor zahlreicher Artikel, Mitglied in verschiedenen Ausschüssen und war bzw. ist als Editor und Gutachter für verschiedene Zeitschriften tätig.

Andreas Meier Andreas Meier ist Professor für Wirtschaftsinformatik an der wirtschafts- und sozialwissenschaftlichen Fakultät der Universität Fribourg, Schweiz. Seine Forschungsgebiete sind eBusiness, eGovernment und Informationsmanagement. Nach Musikstudien in Wien diplomierte er in Mathematik an der ETH in Zürich, doktorierte und habilitierte am Institut für Informatik. Er forschte am IBM Research Lab in Kalifornien/USA, war Systemingenieur bei der IBM Schweiz, Direktor bei der Großbank UBS und Geschäftsleitungsmitglied bei der CSS Versicherung.

Stefan Müller Stefan Müller leitet den Bereich Business Intelligence und Big Data beim Open Source-Spezialisten it-novum GmbH. Nach mehreren Jahren Tätigkeit im Bereich Governance & Controlling sowie Sourcing Management beschäftigt er sich bei it-novum mit dem Einsatz von Open Source Business Intelligence und Big Data Analytics-Lösungen. Er hält regelmäßig Vorträge und publiziert in Fachmedien. Im Dezember 2014 hat er sein erstes Buch zum Thema Pentaho und Jedox veröffentlicht.

Pascal Prassol Pascal Prassol ist als Chief Strategist verantwortlich, SAP in einer Vordenkerrolle zu etablieren und die Geschäftsfeldentwicklung für das Big Data-, IoT- und Platform Services-Business in Deutschland voranzutreiben. Er gestaltet und leitet innovative Transformationsprozesse für Kunden auf Führungsebene und lenkt die Services-Portfoliostrategie mit Fokus auf Innovation in Schlüsselindustrien wie Handel, Konsumgüter, Fertigung und Automobilbau. Pascal Prassol arbeitet seit fast 17 Jahren für SAP in unterschiedlichen nationalen wie internationalen Rollen. Sein Diplom in Wirtschaftsinformatik erlangte er im schönen Saarland.

Thorsten Pröhl Thorsten Pröhl (Dipl.-Phys.) ist wissenschaftlicher Mitarbeiter am Fachgebiet IuK-Management der Technischen Universität Berlin (Prof. Dr. Rüdiger Zarnekow). Seine Forschungsschwerpunkte liegen in den Bereichen IT-Service Management, Cloud Service Management und Big Data. Thorsten Pröhl studierte Physik an der Technischen Universität Berlin mit den Schwerpunkten Festkörperphysik sowie Elektronenmikroskopie und -holografie. Als Zusatzfächer wählte er Informations- und Kommunikationsmanagement sowie Technologie- und Innovationsmanagement. Parallel zum Studium hat er sich mit Fragestellungen aus dem strategischen Einkauf von Logistik-Dienstleistungen auseinandergesetzt, verschiedenartige Praktika bei der Fraunhofer Gesellschaft absolviert, eine webbasierte Branchenlösung für Bilderrahmenkalkulationen entwickelt und eine IT-Unternehmensberatung aufgebaut. Nachdem er selbst erfolgreich Teilnehmer bei Jugend forscht war, engagiert er sich ehrenamtlich als Juror beim Landeswettbewerb Jugend forscht in Berlin. Er ist Autor mehrerer Fachartikel.

Karl Rehrl Karl Rehrl hat 2002 sein Diplomstudium im Fachbereich Informatik an der Johannes Kepler Universität Linz sowie 2011 sein Doktorratsstudium im Fachbereich Geoinformation an der Technischen Universität Wien abgeschlossen. Seit 2004 leitet er bei Salzburg Research die Forschungslinie Mobile und Webbasierte Informationssysteme mit bis zu 15 wissenschaftlichen Mitarbeiterinnen und Mitarbeitern. Seine aktuellen Forschungsschwerpunkte liegen im Bereich der menschlichen Wegfindung, der räumlichen Informationsverarbeitung, der Lokalisierungstechnologien sowie der ortsbasierten Dienste. Karl Rehrl hat 10 Jahre Erfahrung als Projektleiter von Forschungsprojekten, er ist Autor und Co-Autor von mehr als 50 wissenschaftlichen Publikationen und seit 2013 im Editorial Board des International Journals of Location Based Services tätig. 2012 hat Karl Rehrl die Leitung von ITS Austria West, eines von zwei österreichischen Kompetenzzentren für Verkehrstelematik, übernommen und wurde zum Mitglied des ITS Austria Strategic Boards ernannt.

Andreas Ruckstuhl Andreas Ruckstuhl leitet den Schwerpunkt Statistische Datenanalyse am Institut für Datenanalyse und Prozessdesign (IDP) der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW). Seine Forschungs- und Lehrgebiete umfassen statistische und explorative Datenanalyse sowie predictive modelling Methoden. Anwendungserfahrungen hat er unter anderem in den Bereichen Umwelt, predictive Maintenance, Business und Customer Analytics

sowie alternative Investments. Andreas Ruckstuhl hat Mathematik mit Schwerpunkt Statistik an der ETH Zürich studiert, wo er 1995 in angewandter Statistik auch promovierte. Von 1996 bis 1999 war er zugleich Dozent am Department of Statistics and Econometrics und Research Fellow am Centre for Mathematics and Its Applications der Australian National University. 1999 kam er an die ZHAW, wo er sich am Aufbau des IDP und des Bachelor-Studienganges Wirtschaftsingeneurwesen mit seiner analytischen Ausrichtung beteiligte. Er ist Gründungsmitglied des interdisziplinären Labors für Data Science „ZHAW Datalab“ und wirkte als Leiter des CAS Datenanalyse beim Aufbau des neuen ZHAW-Weiterbildungsprogramms Data Science mit.

Heiko Schuldt Heiko Schuldt ist Professor für Informatik am Departement Mathematik und Informatik der Universität Basel. Er studierte Informatik an der Universität Karlsruhe (heute KIT) und war danach wissenschaftlicher Mitarbeiter in der Datenbankgruppe von Prof. Dr. H.-J. Schek an der ETH Zürich, wo er 2000 promovierte. Nach einer Tätigkeit als Oberassistent an der ETH Zürich war Heiko Schuldt von 2003 bis 2006 Professor an der Privaten Universität für Gesundheitswissenschaften, Medizinische Informatik und Technik Tirol (UMIT). Seit Oktober 2005 leitet er an der Universität Basel die Forschungsgruppe Datenbanken und Informationssysteme. Seine Forschungsinteressen umfassen die Datenverwaltung in verteilten Informationssystemen, speziell in der Cloud, Multimedia Retrieval, mobile Informationssysteme und Service-orientierte Architekturen.

Andreas Seufert Andreas Seufert lehrt Betriebswirtschaftslehre und Informationsmanagement an der Hochschule Ludwigshafen und ist Direktor des Instituts für Business Intelligence an der Steinbeis Hochschule Berlin sowie Leiter des Fachkreises „Big Data und Business Intelligence“ des Internationalen Controllervereins. Darüber hinaus ist er als Gutachter renommierter Zeitschriften und Konferenzen tätig. Prof. Dr. Seufert verfügt über langjährige Erfahrung im Bereich der akademischen Forschung und Lehre, u. a. an der Universität St. Gallen sowie dem Massachusetts Institute of Technology (MIT). Als (Co-) Autor und Herausgeber von Büchern, Zeitschriften und Konferenzbeiträgen verfasste er über 100 Publikationen. Er besitzt eine langjährige internationale Erfahrung im Bereich der IT- und Managementberatung. Schwerpunkte seiner internationalen Forschungs- und Beratungstätigkeiten sind Informationsmanagement, Unternehmenssteuerung, Business Intelligence und Big Data.

Thilo Stadelmann Thilo Stadelmann studierte Informatik in Gießen und Marburg aus seiner Leidenschaft für „künstliche Intelligenz“ heraus: Wie lässt sich Maschinen etwas beibringen, das bisher nur Menschen vermochten? Diese Leidenschaft führte ihn über ein Doktorat in Multimedia-Analyse und Maschinellem Lernen sowie Positionen als Datawarehouse- und Applikationsentwickler in das Management eines schwäbischen Mittelständlers. Hier entwickelte er mit seinem Team von 15 hoch qualifizierten IT Architekten und Consultants Datenmanagement- und Data Mining Applikation für die europäische Automobilindustrie. Anfang 2013 kam er

als Dozent für Information Engineering an die Zürcher Hochschule für Angewandte Wissenschaften (ZHAW), an der er eines der ersten interdisziplinären Labore für Data Science in Europa mit gründete: Das ZHAW Datalab, welches er heute leitet. Mit seinen Kollegen vom Datalab entwickelte er eines der ersten Weiterbildungscurricula in Data Science, in denen Professionals in der Kunst und Wissenschaft trainiert werden, Schweizer Unternehmen fit für das Datenzeitalter zu machen.

Kurt Stockinger Kurt Stockinger ist Dozent für Informatik und Studienleiter für Data Science an der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW). Sein Fachgebiet umfasst Data Science mit Fokus auf Big Data, Data Warehousing und Business Intelligence. Er ist auch im Advisory Board der Callista Group AG. Vorher arbeite Kurt Stockinger 5 Jahre als DWH und BI Architekt bei Credit Suisse in Zürich. Er beschäftigte sich mit Design & Implementierungen von Algorithmen für ein unternehmensweites DWH im Terabytebereich, Datensicherheit und Cloud Computing. Stationen seines Werdegangs umfassen 8-jährige Forschungstätigkeiten am Lawrence Berkeley National Laboratory, am CERN und am California Institute of Technology. Er wurde in Informatik am CERN und der Universität Wien promoviert. Kurt Stockinger hat mehr als 60 Publikationen in Journalen und Konferenzen in Wissenschaft und Wirtschaft. Im Jahr 2008 erhielt er einen „R&D 100 Technology Award“ (gilt in den USA als Oscar für Innovationen) gemeinsam mit drei Kollegen vom Berkeley Lab für Forschungsarbeiten an dem Datenbankindex FastBit.

Marcel Wehrle Marcel Wehrle doktoriert aktuell in Wirtschaftsinformatik an der Universität Fribourg, Schweiz. Er hat einen Premaster in Computer Science und einen Master in Kommunikationswissenschaften von der Universität Fribourg, Schweiz. Seine Forschung fokussiert sich auf Granular Computing, Information Granulation, Self Organizing Maps. Nebenberuflich ist er Gesellschafter zweier Unternehmen mit Fokus auf Applikationsentwicklung im Bereich Immobilienbewirtschaftungen.

Rüdiger Zarnekow Rüdiger Zarnekow ist Inhaber des Lehrstuhls für Informations- und Kommunikationsmanagement an der Technischen Universität Berlin. Seine Forschungsschwerpunkte liegen im Bereich des IT-Service-Managements, des strategischen IT Managements und der Geschäftsmodelle für die ICT Industrie. Von 2001 bis 2006 war er am Institut für Wirtschaftsinformatik an der Universität St. Gallen tätig und leitete dort das Competence Center „Industrialisierung im Informationsmanagement“. Prof. Zarnekow promovierte 1999 an der Technischen Universität Freiberg. Von 1995 bis 1998 war er bei der T-Systems Multimedia Solutions GmbH beschäftigt. Er studierte Wirtschaftsinformatik an der European Business School, Oestrich-Winkel und schloss den Master-of-Science in Advanced Software Technologies an der University of Wolverhampton in England ab. Prof. Zarnekow ist freiberuflich als Berater in Fragen des Informationsmanagements und des Electronic Business tätig. Im Jahr 2013 wurde er von der TU Berlin aufgrund seines außerordentlichem Engagement im Bereich Entrepreneurship mit dem Titel

„TUB Entrepreneurship Supporter des Jahres“ ausgezeichnet. Er ist Autor mehrerer Fachbücher und zahlreicher Artikel.

Wolfgang Zimmermann Wolfgang Zimmermann leitet seit Mitte 2013 das Projekt M360 innerhalb des Migros-Genossenschafts-Bund. Mit mehr als 15 Jahren Erfahrung im Online Marketing, CRM und E-Commerce Bereich begleitete er die digitale Transformation seit deren Anfängen und als Diplom Volkswirt und Elektroingenieur bringt er Technik- und Businessverständnis zusammen.

Darius Zumstein Darius Zumstein ist seit 2013 Leiter Digital Analytics & Data Management bei Sanitas Krankenversicherung in Zürich und verantwortet dort das Digital Analytics und (Social) Digital Monitoring. Er studierte Betriebswirtschaftslehre an der Universität Freiburg (Schweiz), wo er am Lehrstuhl für Wirtschaftsinformatik als Forschungsassistent arbeitete und zum Thema Web Analytics promovierte. Als externer Web Analytics Consultant der BMW AG bei FELD M sowie als Web Analytics Manager bei der Scout24 Group und Kabel Deutschland sammelte er wertvolle Berufserfahrung in Deutschland. Seit 2014 leitet er ebenso den Kurs „Online Shop and Sales Management“ an der Hochschule Luzern.

Inhaltsverzeichnis

Teil I

Grundlagen

Was versteht man unter Big Data und NoSQL?

1

Daniel Fasel und Andreas Meier

Zusammenfassung

Verfolgt man die Diskussionen in der europäischen Wirtschaft, erkennt man, dass der Begriff Big Data in der Praxis nicht klar definiert ist. Er ist zwar in aller Munde, doch nur wenige haben eine Antwort auf die Frage, was Big Data ist und welche Unterschiede zu den bestehenden Datenbeständen im Unternehmen existieren. Dieses Kapitel gibt eine Begriffsklärung für Big Data und NoSQL. Anhand der drei Merkmale Volume, Velocity und Variety werden grundlegende Aspekte von Big Data erläutert. Um Big Data wertschöpfend in einer Firma oder Organisation einzusetzen, braucht es Technologien und Fähigkeiten, neben formatierten Daten auch semi-strukturierte und unstrukturierte Daten effizient verarbeiten zu können. Neben den Grundlagen zu SQL- und NoSQL-Datenbanken werden die Kernkompetenzen für ein Datenmanagement im Zeitalter von Big Data aufgezeigt. Weiterführende Literaturangaben runden das Kapitel ab.

Schlüsselwörter

Big Data • Multistrukturelle Daten • Datenanalyse • NoSQL • Datenmanagement

vollständig neuer Original-Beitrag

D. Fasel (✉)
Scigility AG, Zürich, Schweiz
E-Mail: df@scigility.com

A. Meier
Universität Fribourg, Fribourg, Schweiz
E-Mail: andreas.meier@unifr.ch

D. Fasel, A. Meier (Hrsg.), *Big Data*, Edition HMD,
DOI 10.1007/978-3-658-11589-0_1

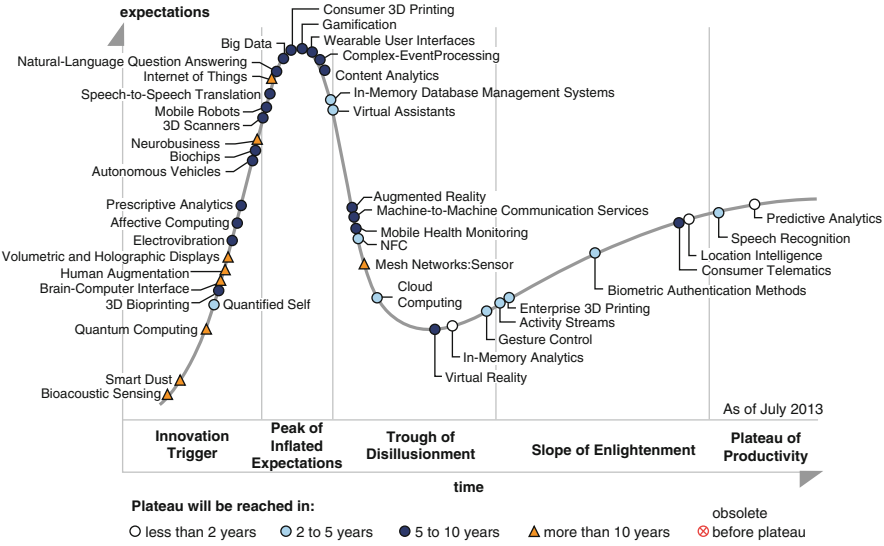


Abb. 1.1 Gartners Hype-Zyklus von 2013 (http://www.gartner.com/newsroom/id/2575515)

1.1 Hype Big Data

Big Data ist ein Schlagwort, das nicht nur in der Fachliteratur Einzug gefunden hat, sondern auch in der Tagespresse breit diskutiert wird: Smart Cities sind ohne die Werkzeuge für Big Data nicht realisierbar, Wettervorhersagen oder Klimaentwicklungen beruhen auf der Analyse umfassender Datenbestände und einzelne Organisationen nutzen die sozialen Netze dazu aus, der Meinungsbildung unterschiedlicher Bedarfsgruppen auf die Spur zu kommen.

Betrachtet man den Hype-Zyklus von Gartner in Abb. 1.1, liegt Big Data beinahe im Zenit der „Inflated Expectations“ (überzogenen Erwartungen).[1] Im europäischen Raum trifft diese Kategorisierung von Gartner nach Erfahrung der Autoren für viele Unternehmen und Organisationen zu.

Unter traditionellen Firmen werden hier vor allem Firmen mit klassischen Strukturen und Geschäftsmodellen verstanden. Im Gegensatz zu den traditionellen Firmen haben webbasierte ihr Geschäftsmodell stark an die Technologieveränderungen im Internet, Cloud Computing sowie Web-Services angepasst. Sie verfügen weitgehend über webbasierte Geschäfts- und Vertriebsmodelle, die Datenhaltung erfolgt in der Cloud resp. auf massiv verteilten Rechnerstrukturen und Webservices ersetzen klassische Transaktionen resp. Geschäftsprozesse.[2]

[1] http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp.

[2] Es ist hier zu vermerken, dass die Autoren die Kategorisierung webbasiert/klassisch nicht trennscharf betrachten. Firmen haben oft ein Mischmodell beider Kategorien und nur wenige sind Reinformen dieser Kategorien.

Als Beispiel von webbasierten Firmen können Amazon, Google oder Facebook genannt werden. Betrachtet man die Herkunft des Begriffs Big Data, ist zu erkennen, dass neben der Forschung vor allem webbasierte Firmen zu den ersten gehörten, die den Begriff prägten. Viele dieser Firmen haben Technologien auf den Markt gebracht, welche heute zu den populärsten Big-Data-Technologien gehören. So präsentierte Google beispielsweise 2003 das Google File System und im Jahr 2006 die Big Table (Chang et al. 2006). 2007 kam Amazon mit Dynamo auf den Markt (DeCandia et al. 2007) und Yahoo hat Hadoop, eine markante Big-Data-Technologie, irgendwann zwischen 2002 und 2006 entwickelt (Harris 2015).[3]

Viele Führungsverantwortliche in Unternehmen und Organisationen sprechen heute von Big Data. Sie weisen in ihren IT-Strategien auf die Vorzüge von Big Data hin und versuchen in Pilotstudien, das Potenzial der entsprechenden Technologien zu erfassen. Einige Unternehmen oder Organisationen haben sogenannte NoSQL-Technologien (siehe Abschn. 1.3.2) mittlerweile im produktiven Einsatz und versuchen, wirtschaftlichen Nutzen aus Big Data zu ziehen.

Vergleicht man die typischen Datenvorkommen webbasierter Firmen und deren Verarbeitung mit Daten traditioneller Firmen, scheint es auf den ersten Blick nicht immer evident zu sein, dass Big-Data-Technologien sinnvoll bei traditionellen Firmen eingesetzt werden können. Dieser Eindruck rührt daher, dass unter Big Data oft ausschließlich große unstrukturierte Daten verstanden werden, welche nicht in relationalen Datenbanken gehalten werden können. Dies ist aber nur ein Teilaspekt von Big Data. Ausserdem sind die meisten Firmen erstaunt, wenn sie entdecken, wie viele geschäftsrelevante Daten sie haben, die genau in die Charakteristik von Big Data passen.

1.2 Definition von Big Data

1.2.1 Volume, Variety, Velocity

Adrian Merv definiert Big Data als Daten, die in ihrer Größe klassische Datenhaltung, Verarbeitung und Analyse auf konventioneller Hardware übersteigen (Merv 2011). Weiterhin beschreibt er Big Data als weitaus heterogener als klassische Daten. So sollen auch externe Daten für analytische Aufgaben in Betracht gezogen werden. Es geht darum, das in sich geschlossene Datenuniversum einer Firma aufzusprengen und mit neuen Daten zu erweitern und damit eine globalere Sicht auf das Unternehmen zu erhalten.

Das McKinsey Global Institute hat 2011 das Phänomen Big Data in verschiedenen Industrien in den USA und Europa untersucht und daraufhin einige Aspekte zusammengetragen, welche sie als Mehrwert von Big Data sehen (Manyika et al. 2011). Big Data erhöht die Informationstransparenz und die Frequenz, wie Daten verarbeitet und analysiert werden können. Durch den größeren Detaillierungsgrad der Daten können erweiterte Applikationen vorangetrieben werden. So lassen sich

[3] Diese Aufzählung ist nicht als vollständig zu betrachten. Es gibt noch eine große Anzahl Firmen, wie Facebook, Twitter, etc., die viel zu diesen Technologien beigetragen haben.

beispielsweise Simulationen auf detaillierten Daten durchführen, die vorher nicht möglich waren. Auch sind Anreicherungen der bestehenden Applikationen möglich, was letztlich zu akkurateren Entscheidungen führen kann, so das McKinsey Global Institute.

Aus den Definitionen von Adrian Merv und dem McKinsey Global Institute lassen sich die folgenden Charakteristiken herauskristallisieren:

- Volume: Der Datenbestand ist umfangreich und liegt im Tera- bis Zettabytebereich (Megabyte = 10^6 Byte, Gigabyte = 10^9 Byte, Terabyte = 10^{12} Byte, Petabyte = 10^{15} Byte, Exabyte = 10^{18} Byte, Zettabyte = 10^{21} Byte).
- Variety: Unter Vielfalt versteht man bei Big Data die Speicherung von strukturierten, semi-strukturierten und unstrukturierten Multimedia-Daten (Text, Grafik, Bilder, Audio und Video).
- Velocity: Der Begriff bedeutet Geschwindigkeit und verlangt, dass Datenströme (Data Streams) in Echtzeit ausgewertet und analysiert werden können.

Die sogenannten 3 V's für Big Data werden von der Gartner Group[4] ebenfalls herausgestrichen. Zudem spricht die Gartner Group von Informationskapital oder Vermögenswert. Aus diesem Grunde fügen einige Experten ein weiteres V zur Definition von Big Data hinzu:

- Value: Big Data Anwendung sollen den Unternehmenswert steigern. Investitionen in Personal und technischer Infrastruktur werden dort gemacht, wo eine Hebelwirkung besteht resp. ein Mehrwert generiert werden kann.

Ein letztes V soll die Diskussion zur Begriffsbildung von Big Data abrunden (vgl. Meier und Kaufmann 2016):

- Veracity: Da viele Daten vage oder ungenau sind, müssen spezifische Algorithmen zur Bewertung der Aussagekraft resp. zur Qualitätseinschätzung der Resultate verwendet werden. Umfangreiche Datenbestände garantieren nicht per se bessere Auswertungsqualität.

Veracity bedeutet in der deutschen Übersetzung Aufrichtigkeit oder Wahrhaftigkeit. Im Zusammenhang mit Big Data wird damit ausgedrückt, dass Datenbestände in unterschiedlicher Datenqualität vorliegen und dass bei Auswertungen dies berücksichtigt werden muss. Neben statistischen Verfahren existieren unscharfe Methoden des Soft Computing, die einem Resultat oder einer Aussage einen Wahrheitswert zwischen ‚richtig' oder ‚falsch' zuordnen (vgl. unscharfe Logik).

1.2.2 Skalieren bei großen Datenmengen

Firmen wie Facebook oder Twitter speichern Petabytes an Daten. Die Struktur eines Facebook-Posts kann reiner Text sein, er kann aber auch ein Bild, ein Video oder einen Link auf einen Inhalt ausserhalb von Facebook enthalten. Webbasierte Firmen

[4] http://www.gartner.com/newsroom/id/1731916.

stießen im letzten Jahrzehnt auf die Probleme, welche Big Data bei herkömmlicher relationaler Datenhaltung verursachen. Sie müssen umfangreiche Datenbestände in kurzer Zeit verarbeiten. Zudem liegen die Daten verteilt über ganze Kontinente hinweg und zusätzlich auf Hardware, die auch mal ausfällt oder unzuverlässig arbeitet.

Relationale Datenbanken stossen bei solchen Aufgabenstellungen an ihre technischen und architektonischen Grenzen. Es wird schwer, Datenvolumen jenseits von 100 Terabyte auf klassischen relationalen Datenbanken zu speichern. Wenn man herkömmliche Datenbanken auf mehrere physische Maschinen verteilt, erhöht sich die Komplexität, die Datenbank operativ stabil und die Daten konsistent zu halten. Auch die Hardwarekosten steigen entsprechend an, da hochperformante und adaptierte Systeme benötigt werden. Um diese Probleme zu entschärfen, setzen webbasierte Firmen NoSQL-Technologien ein, die horizontal skalieren und auf günstigerer Mainstream-Hardware betrieben werden.

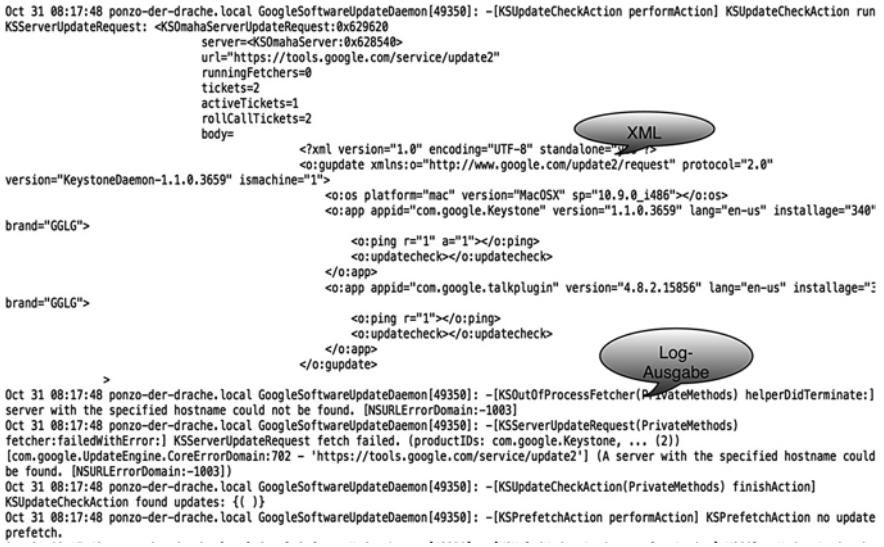
Der Begriff horizontales Skalieren bedeutet: Will man mehr Leistung, muss man einfach mehr Maschinen in den Systemverbund[5] einschließen. Im Gegensatz dazu müsste man beim vertikalen Skalieren die Leistungsfähigkeit der einzelnen Maschinen ausbauen, beispielsweise bessere CPU's oder mehr Arbeitsspeicher einbauen. Vertikales Skalieren führt dazu, dass die Maschinen schnell teuer werden, da die besten und neuesten Einzelkomponenten zum Einsatz kommen. Beim horizontalen Skalieren kann man ältere und billigere Komponenten einsetzen und über die Menge der Maschinen die Leistungsfähigkeit ausbauen. Bei Clouds, wo man binnen weniger Minuten neue Maschinen in einen Verbund einbinden und dann auch wieder entfernen kann, benutzt man oft den Begriff elastisch, um diese Art von horizontalem Skalieren zu umschreiben.

1.2.3 Verarbeitung multi-struktureller Daten

Neben dem Abspeichern von großen Volumen in hochverteilten Systemen ist die Vielfalt der Strukturen eine weitere Schwierigkeit, welche nur bedingt mit klassischen Datenbanken angegangen werden kann. In relationalen Datenbanken können zwar binäre Datenformate mittels dem Datentyp Binary Large Objects (BLOB) gespeichert werden, jedoch ist diese Form von Datenhaltung meist ungeeignet für größere Dateien. Die Restriktion relationaler Datenbanken, alle Datenvorkommen in Tabellen zu fassen und jede Transaktion konsistent zu halten, ist nicht immer möglich und wünschenswert (Fasel 2014). So können Daten von externen Quellen, wie beispielsweise Webservices, ihre Formate flexibel gestalten und schnell ändern, teils sogar bei jeder Anfrage.

Ein weiteres Beispiel von multistrukturellen Daten sind klassische Log-Dateien von Servern. Sie zeigen in sich verschachtelte Strukturen, die sich je nach Art des Log-Eintrages unterscheiden. Abb. 1.2 zeigt zur Visualisierung einen Auszug aus einer Log-Datei von einem Apple MacBook Pro.

[5] Systemverbund ist der Versuch der Autoren das englische Wort „Cluster" in diesem Kontext zu übersetzen. Im Folgenden werden die Wörter Cluster und Systemverbund synonym verwendet.




```
Oct 31 08:17:48 ponzo-der-drache.local GoogleSoftwareUpdateDaemon[49350]: -[KSUpdateCheckAction performAction] KSUpdateCheckAction run
KSServerUpdateRequest: <KSOmahaServerUpdateRequest:0x629620
                        server=<KSOmahaServer:0x628540>
                        url="https://tools.google.com/service/update2"
                        runningFetchers=0
                        tickets=2
                        activeTickets=1
                        rollCallTickets=2
                        body=
                                <?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
                                <o:gupdate xmlns:o="http://www.google.com/update2/request" protocol="2.0"
version="KeystoneDaemon-1.1.0.3659" ismachine="1">
                                    <o:os platform="mac" version="MacOSX" sp="10.9.0_i486"></o:os>
                                    <o:app appid="com.google.Keystone" version="1.1.0.3659" lang="en-us" installage="340"
brand="GGLG">
                                        <o:ping r="1" a="1"></o:ping>
                                        <o:updatecheck></o:updatecheck>
                                    </o:app>
                                    <o:app appid="com.google.talkplugin" version="4.8.2.15856" lang="en-us" installage="3
brand="GGLG">
                                        <o:ping r="1"></o:ping>
                                        <o:updatecheck></o:updatecheck>
                                    </o:app>
                                </o:gupdate>
            >
Oct 31 08:17:48 ponzo-der-drache.local GoogleSoftwareUpdateDaemon[49350]: -[KSOutOfProcessFetcher(PrivateMethods) helperDidTerminate:]
server with the specified hostname could not be found. [NSURLErrorDomain:-1003]
Oct 31 08:17:48 ponzo-der-drache.local GoogleSoftwareUpdateDaemon[49350]: -[KSServerUpdateRequest(PrivateMethods)
fetcher:failedWithError:] KSServerUpdateRequest fetch failed. (productIDs: com.google.Keystone, ... (2))
(com.google.UpdateEngine.CoreErrorDomain:702 - 'https://tools.google.com/service/update2'] (A server with the specified hostname could
be found. [NSURLErrorDomain:-1003])
Oct 31 08:17:48 ponzo-der-drache.local GoogleSoftwareUpdateDaemon[49350]: -[KSUpdateCheckAction(PrivateMethods) finishAction]
KSUpdateCheckAction found updates: {( )}
Oct 31 08:17:48 ponzo-der-drache.local GoogleSoftwareUpdateDaemon[49350]: -[KSPrefetchAction performAction] KSPrefetchAction no update
prefetch.
```

Abb. 1.2 Multistrukturelle Log-Datei eines Apple MacBook Pro aus Fasel 2014

Wie zu erkennen ist, mischt die Log-Datei XML-Formate und einfache Log-Ausgaben. Solche Dateien erhöhen die Komplexität, ein relationales Datenbankschema zu definieren. Wenn machbar, führt es auf jeden Fall zu großen Konzeptions- und Wartungsaufwänden.

1.2.4 Analyse verteilter Daten

Bei Amazon fallen täglich Millionen von Transaktionen aus allen Teilen der Welt in ihrem Webshop an. In einem 2007 veröffentlichten Artikel über Amazons Dynamo (DeCandia et al. 2007) schreiben die Autoren von über 10 Millionen Transaktionen und über 3 Millionen Checkouts (Einkäufe) pro Tag. Um diese Masse an Transaktionen verteilt zu bewältigen, wurde die NoSQL-Datenbank Dynamo entwickelt. Diese ist verteilt und hochverfügbar, aber nicht immer konsistent. Das ist ein direktes Antiparadigma zu klassischen relationalen Datenbanken. Wenn ein Warenkorb inkonsistente Daten von Dynamo erhält, entscheidet der Warenkorb selbst, wie er damit umgehen soll. Der Artikel zeigt auch, dass Amazon mit Dynamo über 99.9995 % Verfügbarkeit weltweit garantieren kann und nur 0.06 % aller Daten inkonsistent an die Warenkörbe geliefert werden.

Hand in Hand mit den NoSQL-Technologien, die effizienteres Speichern und Analysieren ermöglichen, etablieren sich Techniken, wie man große und multistrukturelle Daten besser analysieren kann (Fasel 2014). Technologien, die horizontal skalieren, verteilen die Daten auf die partizipierenden Maschinen im Cluster. Durch diese Verteilung der Daten lassen sich Analysen auf Sub-Sets parallel durchführen. So kann zuerst die Datenpartition auf einem lokalen Knoten im Cluster

verarbeitet werden, in einem zweiten Schritt werden die Teilresultate zusammengefasst (vgl. Map/Reduce-Verfahren).

Der Vorteile einer verteilten Berechnung besteht in der Parallelisierung und der Datenlokalität. Die Daten werden nicht zur berechnenden Entität gebracht, sondern die Berechnungsvorschrift wird zu den einzelnen Teildaten gebracht. Solche verteilten Berechnungen haben aber auch Nachteile. Beispielsweise lassen sich einige mathematische Funktionen nicht auf einzelnen Teilbereichen berechnen und anschließend auf das gesamte Datenset aggregieren. Ein Median beispielsweise kann nicht auf zehn Sub-Sets einzeln berechnet und daraus der globale Median aus den Teilmedianen hergeleitet werden.

Eine weitere Schwierigkeit bei Analysen von großen, unstrukturierten und heterogenen Datenmengen liegt darin, dass Scheinkorrelationen entstehen können (Fasel 2014). Betrachtet man beispielsweise die Schneeschmelze der Gletscher der Schweizer Alpen in den letzten hundert Jahren und die Entwicklung des Immobilienmarktes an der Zürcher Goldküste, könnte es gut sein, dass eine Korrelation erkannt werden kann. Die Kausalität dieser Korrelation ist aber für die Bestimmung des heutigen Marktwertes eines Grundstückes an der Goldküste irrelevant. Scheinkorrelationen können bei großen und heterogenen Datensets aus purem Zufall entstehen. Bill Franks (Franks 2012) geht explizit auf diese Problematik von Big Data in seinem Buch ein. Er beschreibt große Sets von Daten aus unterschiedlichen Quellen als chaotisch und hässlich. Mit dieser Aussage spricht er die unterschiedliche Qualität der Daten, die Problematik, die verschiedenen Quellen vernünftig zu kombinieren, und auch die Gefahr von Scheinkorrelationen an.

Bill Franks leitet aus dieser Problematik das bereits erwähnte vierte V für Big Data ab: Value. Value steht hier für die Charakteristik von Big Data, einen Mehrwert für das Unternehmen zu bieten und somit die Daten für zielgerichtete und wertschöpfende Geschäftsfälle zu verwenden.

Um die Problematiken zu entschärfen, können moderne Techniken, wie selbstlernende Algorithmen, Clustering-Methoden oder auch Algorithmen zum Erstellen von Prognosen verwendet werden. Diese müssen aber für verteilte Systeme optimiert sein; es eignen sich nicht alle Techniken gleich, um eine spezifische Problemstellung lösen zu können.

1.3 SQL- und NoSQL-Technologien

1.3.1 Relationale Datenbanken

Das Relationenmodell wurde Anfang der siebziger Jahre des letzten Jahrhunderts durch die Arbeiten von Edgar Frank Codd begründet. Daraufhin entstanden in Forschungslabors erste relationale Datenbanksysteme, die SQL (Structured Query Language) oder ähnliche Datenbanksprachen unterstützten. Ausgereiftere Produkte haben inzwischen die Praxis erobert.

Ein relationales Datenbanksystem ist gemäß Abb. 1.3 ein integriertes System zur einheitlichen Verwaltung von Tabellen. Neben Dienstfunktionen stellt es die

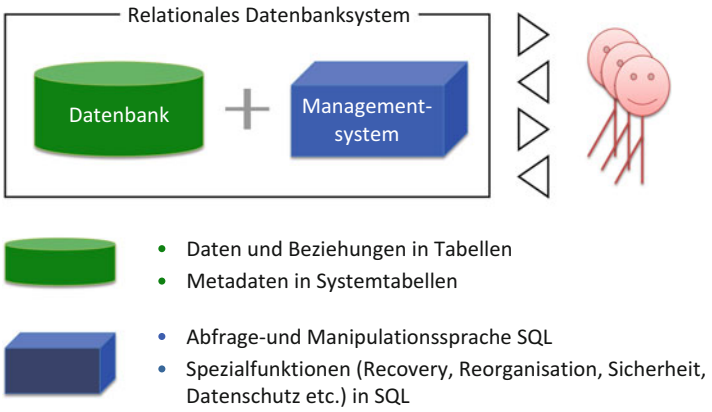


Abb. 1.3 Die zwei Komponenten eines relationalen Datenbanksystems

deskriptive Sprache SQL für Datenbeschreibungen, Datenmanipulationen und -selektionen zur Verfügung.

Jedes relationale Datenbanksystem besteht aus einer Speicherungs- und einer Verwaltungskomponente: Die Speicherungskomponente dient dazu, sowohl Daten als auch Beziehungen zwischen ihnen lückenlos in Tabellen abzulegen. Neben Tabellen mit Benutzerdaten aus unterschiedlichen Anwendungen existieren vordefinierte Systemtabellen, die beim Betrieb der Datenbanken benötigt werden. Diese enthalten Beschreibungsinformationen und lassen sich vom Anwender jederzeit abfragen, nicht aber verändern.

Die Verwaltungskomponente enthält als wichtigsten Bestandteil die relationale Datendefinitions-, Datenselektions- und Datenmanipulationssprache SQL. Daneben umfasst diese Sprache auch Dienstfunktionen für die Wiederherstellung von Datenbeständen nach einem Fehlerfall, zum Datenschutz und zur Datensicherung.

Die Eigenschaften eines relationalen Datenbanksystems lassen sich wie folgt zusammenfassen (vgl. Meier und Kaufmann 2016):

- Modell: Das Datenmodell ist relational, d. h. alle Daten werden in Tabellen abgelegt. Abhängigkeiten zwischen den Merkmalswerten einer Tabelle oder mehrfach vorkommende Sachverhalte können aufgedeckt werden. Die dazu notwendigen formalen Instrumente (sog. Normalformen) ermöglichen einen widerspruchsfreien Datenbankentwurf und garantieren saubere Datenstrukturen.
- Architektur: Das System gewährleistet eine große Datenunabhängigkeit, d. h. Daten und Anwendungsprogramme bleiben weitgehend voneinander getrennt. Diese Unabhängigkeit ergibt sich aus der Tatsache, dass die eigentliche Speicherungskomponente von der Anwenderseite durch eine Verwaltungskomponente entkoppelt ist. Im Idealfall können physische Änderungen in den relationalen Datenbanken vorgenommen werden, ohne dass die entsprechenden Anwendungsprogramme anzupassen sind.

- Schema: Die Definition der Tabellen und der Merkmale werden im relationalen Datenbankschema abgelegt. Dieses enthält zudem die Definition der Identifikationsschlüssel sowie Regeln zur Gewährung der Integrität.
- Sprache: Das Datenbanksystem verwendet SQL zur Datendefinition, -selektion und -manipulation. Die Sprachkomponente ist deskriptiv und entlastet den Anwender bei Auswertungen oder bei Programmiertätigkeiten.
- Mehrbenutzerbetrieb: Das System unterstützt den Mehrbenutzerbetrieb, d. h. es können mehrere Benutzer gleichzeitig ein und dieselbe Datenbank abfragen oder bearbeiten. Das relationale Datenbanksystem sorgt dafür, dass parallel ablaufende Transaktionen auf einer Datenbank sich nicht gegenseitig behindern oder gar die Korrektheit der Daten beeinträchtigen.
- Konsistenzgewährung: Ein relationales Datenbanksystem stellt Hilfsmittel zur Gewährleistung der Datenintegrität bereit. Unter Datenintegrität versteht man die fehlerfreie und korrekte Speicherung der Daten sowie ihren Schutz vor Zerstörung, vor Verlust, vor unbefugtem Zugriff und Missbrauch.

Nicht-relationale Datenbanksysteme erfüllen obige Eigenschaften nur teilweise. Aus diesem Grunde sind die relationalen Datenbanksysteme in den meisten Unternehmen, Organisationen und vor allem in KMU's (kleinere und mittlere Unternehmen) nicht mehr wegzudenken. Zudem legen relationale Datenbanksysteme auf dem Gebiet der Leistungsfähigkeit von Jahr zu Jahr zu, obwohl mengenorientierte Verarbeitung und Konsistenzsicherung ihren Preis haben. Bei massiv verteilten Anwendungen im Web hingegen oder bei Big-Data-Anwendungen muss die relationale Datenbanktechnologie mit NoSQL-Technologien ergänzt werden, um Webdienste rund um die Uhr und weltweit anbieten zu können.

1.3.2 NoSQL-Datenbanken

Nicht-relationale Datenbanken gab es vor der Entdeckung des Relationenmodells durch Ted Codd in der Form von hierarchischen oder netzwerkartigen Datenbanken. Nach dem Aufkommen von relationalen Datenbanksystemen wurden nicht-relationale Ansätze weiterhin für technische oder wissenschaftliche Anwendungen genutzt. Beispielsweise war es schwierig, ein CAD-System (CAD = Computer Aided Design) für Bau- oder Maschinenteile mit relationaler Technologie zu betreiben. Das Aufteilen technischer Objekte in eine Vielzahl von Tabellen war für CAD-Systeme problematisch, da geometrische, topologische und grafische Manipulationen in Echtzeit durchgeführt werden mussten (Meier 1987).

Mit dem Aufkommen des Internet und einer Vielzahl von webbasierten Anwendungen haben nicht-relationale Datenkonzepte gegenüber relationalen an Gewicht gewonnen. Es ist schwierig oder teilweise unmöglich, Big-Data-Anwendungen mit relationaler Datenbanktechnologie zu bewältigen.

Die Bezeichnung ‚nicht-relational' wäre besser geeignet als NoSQL, doch hat sich der Begriff in den letzten Jahren bei Datenbankforschern wie bei Anbietern im Markt etabliert. Der Begriff NoSQL wird heute für nicht-relationale Ansätze im

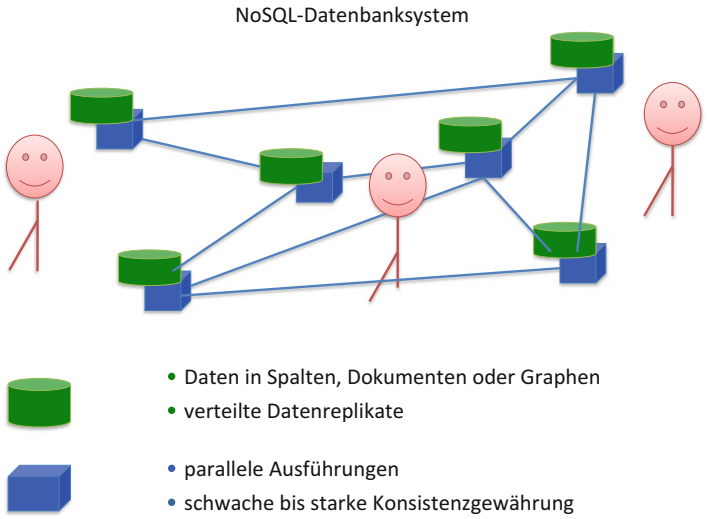


Abb. 1.4 Grundstruktur eines NoSQL-Datenbanksystems

Datenmanagement verwendet. Manchmal wird der Ausdruck NoSQL durch ‚Not only SQL' übersetzt. Damit soll ausgedrückt werden, dass bei einer massiv verteilten Webanwendung nicht nur relationale Datentechnologien zum Einsatz gelangen. Vor allem dort, wo die Verfügbarkeit des Webdienstes im Vordergrund steht, sind NoSQL-Technologien gefragt.

Ein NoSQL-Datenbanksystem unterliegt gemäß Abb. 1.4 einer massiv verteilten Datenhaltungsarchitektur. Die Daten selber werden je nach Typ der NoSQL-Datenbank entweder als Schlüssel-Wertpaare (Key/Value Store), in Spalten oder Spaltenfamilien (Column Store), in Dokumentspeichern (Document Store) oder in Graphen (Graph Database) gehalten (siehe Kap. 2 und 6). Um hohe Verfügbarkeit zu gewähren und das NoSQL-Datenbanksystem gegen Ausfälle zu schützen, werden unterschiedliche Replikationskonzepte unterstützt (vgl. z. B. das Konzept Consistent Hashing in Meier und Kaufmann 2016).

Bei einer massiv verteilten und replizierten Rechnerarchitektur können parallele Auswertungsverfahren genutzt werden (vgl. Map/Reduce in Kap. 6). Die Analyse umfangreicher Datenvolumen oder das Suchen nach bestimmten Sachverhalten kann mit verteilten Berechnungsvorgängen beschleunigt werden. Beim Map/Reduce-Verfahren werden Teilaufgaben an diverse Rechnerknoten verteilt und einfache Schlüssel-Wertpaare extrahiert (Map) bevor die Teilresultate zusammengefasst und ausgegeben werden (Reduce).

In massiv verteilten Rechnernetzen werden zudem differenzierte Konsistenzkonzepte angeboten (vgl. Kap. 2). Unter starker Konsistenz (strong consistency) wird verstanden, dass das NoSQL-Datenbanksystem die Konsistenz jederzeit gewährleistet. Bei der schwachen Konsistenzforderung (weak consistency) wird toleriert, dass Änderungen auf replizierten Knoten verzögert durchgeführt und zu kurzfristigen Inkonsistenzen führen können. Daneben existieren weitere

Differenzierungsoptionen, wie z. B. Consistency by Quorum (vgl. Meier und Kaufmann 2016).

Die folgende Definition für NoSQL-Datenbanken ist angelehnt an das webbasierte NoSQL-Archiv[6] sowie an das Textbuch von Meier und Kaufmann 2016. Webbasierte Speichersysteme werden demnach als NoSQL-Datenbanksysteme bezeichnet, falls sie folgende Bedingungen erfüllen:

- Modell: Das zugrunde liegende Datenmodell ist nicht relational.
- Architektur: Die Datenarchitektur unterstützt massiv verteilte Webanwendungen und horizontale Skalierung.
- Mindestens 3V: Das Datenbanksystem erfüllt die Anforderungen für umfangreiche Datenbestände (Volume), flexible Datenstrukturen (Variety) und Echtzeitverarbeitung (Velocity).
- Schema: Das Datenbanksystem unterliegt keinem fixen Datenbankschema.
- Replikation: Das Datenbanksystem unterstützt die Datenreplikation.
- Mehrbenutzerbetrieb: Der Mehrbenutzerbetrieb wird unterstützt, wobei differenzierte Konsistenzeinstellungen gewählt werden können.
- Konsistenzgewährung: Aufgrund des CAP-Theorems (vgl. Kap. 2) ist die Konsistenz lediglich verzögert gewährleistet (weak consistency), falls hohe Verfügbarkeit und Ausfalltoleranz angestrebt werden.

Die Forscher und Betreiber des NoSQL-Archivs listen auf ihrer Webplattform 150 NoSQL-Datenbankprodukte. Der Großteil dieser Systeme ist Open Source. Allerdings zeigt die Vielfalt der Angebote auf, dass der Markt von NoSQL-Lösungen noch unsicher ist. Zudem müssen für den Einsatz von geeigneten NoSQL-Technologien Spezialisten gefunden werden, die nicht nur die Konzepte beherrschen, sondern auch die vielfältigen Architekturansätze und Werkzeuge (vgl. Abschn. 1.5).

1.4 Organisation des Datenmanagements

Viele Firmen und Institutionen betrachten ihre Datenbestände als unentbehrliche Ressource. Sie pflegen und unterhalten zu Geschäftszwecken nicht nur ihre eigenen Daten, sondern schließen sich mehr und mehr an öffentlich zugängliche Datensammlungen an. Das stetige Wachstum der Informationsanbieter mit ihren Dienstleistungen rund um die Uhr untermauert den Stellenwert webbasierter Datenbestände.

Die Bedeutung aktueller und realitätsbezogener Information hat einen direkten Einfluss auf die Ausgestaltung des Informatikbereiches. So sind vielerorts Stellen des Datenmanagements entstanden, um die datenbezogenen Aufgaben und Pflichten bewusster angehen zu können. Ein zukunftsgerichtetes Datenmanagement befasst sich sowohl strategisch mit der Informationsbeschaffung und -bewirtschaftung als auch operativ mit der effizienten Bereitstellung und Auswertung von aktuellen und konsistenten Daten.

[6] NoSQL-Archiv; http://nosql-database.org/, Zugegriffen am 17.02.2015.

Aufbau und Betrieb eines Datenmanagements verursachen Kosten mit anfänglich nur schwer messbarem Nutzen. Es ist nicht immer einfach, eine flexible Datenarchitektur, widerspruchsfreie und für jedermann verständliche Datenbeschreibungen, saubere und konsistente Datenbestände, griffige Sicherheitskonzepte, aktuelle Auskunftsbereitschaft und anderes mehr eindeutig zu bewerten und aussagekräftig in Wirtschaftlichkeitsüberlegungen einzubeziehen. Erst ein allmähliches Bewusstwerden von Bedeutung und Langlebigkeit der Daten relativiert für das Unternehmen die notwendigen Investitionen.

Um den Begriff Datenmanagement besser fassen zu können, sollte das Datenmanagement in seine Aufgabenbereiche Datenarchitektur, Datentechnik und Datennutzung aufgegliedert werden:

- Datenarchitektur: Neben der eigentlichen Analyse der Daten- und Informationsbedürfnisse müssen die wichtigsten Datenklassen und ihre gegenseitigen Beziehungen untereinander in unterschiedlichster Detaillierung analysiert und modelliert werden (vgl. das Relationen- resp. Graphenmodell in Kap. 2).
- Datentechnik: Die Spezialisten der Datentechnik installieren, überwachen und reorganisieren SQL- und NoSQL-Datenbanken und stellen diese in einem mehrstufigen Verfahren sicher.
- Datennutzung: Mit einem besonderen Team von Datenspezialisten (Berufsbild Data Scientist, siehe unten resp. Kap. 4) wird das Business Analytics vorangetrieben, das der Geschäftsleitung und dem Management periodisch Datenanalysen erarbeitet und rapportiert. Zudem unterstützen diese Spezialisten diverse Fachabteilungen wie Marketing, Verkauf, Kundendienst etc., um spezifische Erkenntnisse aus Big Data zu generieren.

Für das Datenmanagement sind im Laufe der Jahre unterschiedliche Berufsbilder entstanden. Die wichtigsten lauten:

- Datenarchitekt: Datenarchitekten sind für die unternehmensweite Datenarchitektur verantwortlich. Aufgrund der Geschäftsmodelle entscheiden sie, wo und in welcher Form Datenbestände bereitgestellt werden müssen. Für die Fragen der Verteilung, Replikation oder Fragmentierung der Daten arbeiten sie mit den Datenbankspezialisten zusammen.
- Datenbankspezialist: Die Datenbankspezialisten beherrschen die Datenbank- und Systemtechnik und sind für die physische Auslegung der Datenarchitektur verantwortlich. Sie entscheiden, welche Datenbanksysteme (SQL- und/oder NoSQL-Technologien) für welche Komponenten der Anwendungsarchitektur eingesetzt werden. Zudem legen sie das Verteilungskonzept fest und sind zuständig für die Archivierung, Reorganisation und Restaurierung der Datenbestände.
- Data Scientist: Die Data Scientists sind die Spezialisten des Business Analytics. Sie beschäftigen sich mit der Datenanalyse und -interpretation, extrahieren noch nicht bekannte Fakten aus den Daten (Wissensgenerierung) und erstellen bei Bedarf Zukunftsprognosen über die Geschäftsentwicklung. Sie beherrschen die Methoden und Werkzeuge des Data Mining (Mustererkennung), der Statistik und der Visualisierung von mehrdimensionalen Zusammenhängen unter den Daten.

Die hier vorgeschlagene Begriffsbildung zum Datenmanagement sowie zu den Berufsbildern umfasst technische, organisatorische wie betriebliche Funktionen. Dies bedeutet allerdings nicht zwangsläufig, dass in der Aufbauorganisation eines Unternehmens oder Organisation die Funktionen der Datenarchitektur, der Datentechnik und der Datennutzung in einer einzigen Organisationseinheit zusammengezogen werden müssen.

1.5 Weiterführende Literaturangaben

Dieses Kapitel beruht auf dem Überblicksbeitrag ‚Big Data – Eine Einführung' von Daniel Fasel (2014) sowie aus Auszügen aus dem Textbuch ‚SQL- & NoSQL-Datenbanken' von Meier und Kaufmann (2016).

Was Big Data betrifft, so ist der Markt in den letzten Jahren mit Büchern überschwemmt worden. Allerdings beschreiben die meisten Werke den Trend nur oberflächlich. Zwei englische Kurzeinführungen, das Buch von Celko (2014) sowie dasjenige von Sadalage und Fowler (2013), erläutern die Begriffe und stellen die wichtigsten NoSQL-Datenbankansätze vor. Für technisch Interessierte gibt es das Werk von Redmond und Wilson (2012), die sieben Datenbanksysteme konkret erläutern.

Erste deutschsprachige Veröffentlichungen zum Themengebiet Big Data gibt es ebenfalls: Das Textbuch von Meier und Kaufmann (2016) zeigt sowohl die Grundlagen für SQL- wie für NoSQL-Datenbanken: Modellierungsaspekte mit Tabellen resp. mit Graphen, relationale und graphorientierte Abfrage- und Manipulationssprachen, Konsistenzbetrachtungen (CAP-Theorem, ACID und BASE, Vektoruhren etc.), Systemarchitektur sowie eine Übersicht über postrelationale Datenbanken (objekt-relationale, föderierte, temporale, multidimensionale und wissensbasierte Datenbanken sowie Fuzzy-Datenbanken) und NoSQL-Datenbanken (Key/Value Store, Column Store, Document Store, XML-Datenbanken, Graphdatenbanken). Das Buch von Edlich et al. (2011) gibt eine Einführung in NoSQL-Datenbanktechnologien, bevor unterschiedliche Datenbankprodukte für Key/Value Store, Document Store, Column Store und Graphdatenbanken vorgestellt werden. Das Werk von Freiknecht (2014) beschreibt das bekannte System Hadoop (Framework für skalierbare und verteilte Systeme) inkl. der Komponenten für die Datenhaltung (HBase) und für das Data Warehousing (Hive). Das HMD-Schwerpunktheft über ‚Big Data' von Fasel und Meier (2014) gibt einen Überblick über die Big-Data-Entwicklung im betrieblichen Umfeld. Die wichtigsten NoSQL-Datenbanken werden vorgestellt, Fallbeispiele diskutiert, rechtliche Aspekte erläutert und Umsetzungshinweise gegeben.

Literatur

Celko, J.: Joe Celko's Complete Guide to NoSQL – What Every SQL Professional Needs to Know About Nonrelational Databases. Morgan Kaufmann, Waltham (2014)

Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., Gruber, R.E.: Bigtable – A distributed storage system for structured data. In: Seventh USENIX

Symposium on Operating System Design and Implementation, OSDI'2006, Seattle, 6–8 Nov (2006)
De Candia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., Vogels, W.: Dynamo – Amazon's Highly Available Key-value Store. 21st ACM Symposium on Operating Systems Principles, SOSP'07, S. 205–230, Stevenson, 14–17 Oct (2007)
Edlich, S., Friedland, A., Hampe, J., Brauer, B., Brückner, M.: NoSQL – Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. Carl Hanser Verlag, München (2011)
Fasel, D.: Big Data – Eine Einführung. In: Fasel, D., Meier, A. (Hrsg.) Big Data. HMD-Zeitschrift Praxis der Wirtschaftsinformatik, Nr. 298, S. 386–400. Springer, Heidelberg (2014)
Fasel, D., Meier, A. (Hrsg.): Big Data. HMD-Zeitschrift Praxis der Wirtschaftsinformatik, Nr. 298. Springer, Heidelberg (2014)
Franks, B.: Taming the Big Data Tidal Wave. Wiley, Heidelberg (2012)
Freiknecht, J.: Big Data in der Praxis – Lösungen mit Hadoop, HBase und Hive – Daten speichern, aufbereiten und visualisieren. Carl Hanser Verlag, München (2014)
Harris, D.: The history of hadoop: From 4 nodes to the future of data. http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/ (2015). Zugegriffen im März (2015)
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, Ch., Byers, A.H.: Big Data – The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute (2011)
Meier, A.: Erweiterung relationaler Datenbanksysteme für technische Anwendungen. Informatik-Fachberichte, Nr. 135, Springer, Berlin (1987)
Meier, A., Kaufmann, M.: SQL- & NoSQL-Datenbanken. Springer (2016)
Merv, A.: It's going mainstream, and it's your next opportunity. Teradata Magazine, 01, (2011)
Redmond, E., Wilson, J.R.: Seven Databases in Seven Weeks – A Guide to Modern Databases and the NoSQL Movement. The Pragmatic Bookshelf, Dallas (2012)
Sadalage, P.J., Fowler, M.: NoSQL Distilled – A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley, Upper Saddle River (2013)

Datenmanagement mit SQL und NoSQL

2

Andreas Meier

Zusammenfassung

Viele webbasierte Anwendungen setzen für die unterschiedlichen Dienste adäquate Datenhaltungssysteme ein. Die Nutzung einer einzigen Datenbanktechnologie, z. B. der relationalen, genügt nicht mehr. In diesem Kapitel werden entsprechend die Grundlagen für relationale Datenbanken – SQL-Datenbanken – sowie für NoSQL-Datenbanken gegeben. Als Einstieg dient ein elektronischer Shop, welcher gleichzeitig SQL- und NoSQL-Datenbanken als Architekturkomponenten beansprucht. Danach werden Modellierungsansätze für den Einsatz von relationalen und graphorientierten Datenbanken einander gegenüber gestellt. Die Nutzung von Daten mittels Datenbankabfragesprachen wird exemplarisch mit SQL (Structured Query Language) für relationale und mit Cypher für graphorientierte Datenbanken illustriert. Zudem werden unterschiedliche Konsistenzvarianten besprochen.

Schlüsselwörter

Semantische Modellbildung • Relationenmodell • Graphenmodell • Abfragesprachen • Konsistenz

Dieses Kapitel beruht teilweise auf Auszügen aus den beiden Textbüchern ‚Relationale und postrelationale Datenbanken' (Meier 2010) resp. ‚SQL- und NoSQL-Datenbanken' (Meier und Kaufmann 2016).

A. Meier (✉)
Universität Fribourg, Fribourg, Schweiz
E-Mail: andreas.meier@unifr.ch

D. Fasel, A. Meier (Hrsg.), *Big Data*, Edition HMD,
DOI 10.1007/978-3-658-11589-0_2

2.1 Nutzung strukturierter und unstrukturierter Daten

Mit dem Aufkommen der Computer wurden die Daten bald auf Sekundärspeichern wie Band, Magnettrommel oder Magnetplatte gehalten. Das Merkmal solcher Datenhaltungssysteme war der wahlfreie oder direkte Zugriff auf das Speichermedium. Die Daten waren strukturiert und mit der Hilfe einer Adresse konnte ein bestimmter Datensatz selektiert werden, meistens unter Nutzung eines Index oder einer Hash-Funktion.

Die Großrechner mit ihren Dateisystemen wurden zu Beginn für technisch-wissenschaftliche Anwendungen genutzt. Der Rechner war ein Zahlenkalkulator resp. Computer im eigentlichen Sinne, ohne den z. B. das Projekt ‚Man on the Moon' keine Chance zum Erfolg gehabt hätte. Mit dem Aufkommen von Datenbanksystemen (CODASYL, hierarchische, relationale) eroberten die Rechner die Wirtschaft: Der Computer wurde zum Zahlen- und Wortkalkulator (vgl. Gugerli et al. 2014). Rechner mit Datenbanksystemen entwickelten sich zum Rückgrat administrativer und kommerzieller Anwendungen, da ein Mehrbenutzerbetrieb auf konsistente Art und Weise unterstützt werden konnte (vgl. ACID in Abschn. 2.4.1).

Nach wie vor basieren die meisten Informationssysteme in Organisationen und Unternehmen auf der relationalen Datenbanktechnik, welche die früher eingesetzten hierarchischen oder netzwerkartigen Datenbanksysteme ablöste. Relationale Systeme verarbeiten strukturierte und formatierte Daten in Form von Tabellen. Zudem muss die Struktur der Daten inkl. der verwendeten Datentypen dem Datenbanksystem durch die Spezifikation eines Schemas mitgeteilt werden (vgl. CREATE TABLE Befehl von SQL). Bei jedem SQL-Aufruf werden die Systemtabellen konsultiert, um u. a. Autorisierungs- und Datenschutzbestimmungen zu prüfen. Erweiterungen von SQL lassen es zu, Buchstabenfolgen (CHARACTER VARYING), Bitfolgen (BIT VARYING, BINARY LARGE OBJECT) oder Textstücke (CHARACTER LARGE OBJECT) zu verarbeiten. Zudem wird die Einbindung von XML (eXtensible Markup Language) unterstützt (Meier und Kaufmann 2016).

Mit dem Aufkommen des Webs resp. webbasierter Dienste haben sich neben der relationalen Datenbanktechnik vor allem NoSQL-Ansätze bewährt. Diese können strukturierte, semi-strukturierte und unstrukturierte Daten sowie Datenströme (Multimedia) in Echtzeit verarbeiten. Es lassen sich differenzierte Verfahren zur Konsistenzgewährung, Verfügbarkeit und Ausfalltoleranz nicht zuletzt aufgrund des sogenannten CAP-Theorems (Consistency, Availability, Partition Tolerance; vgl. Abschn. 2.4.2) mit Einschränkungen kombinieren.

In Abb. 2.1 ist ein elektronischer Shop schematisch dargestellt. Um eine hohe Verfügbarkeit und Ausfalltoleranz zu garantieren, wird ein Key/Value-Speichersystem (siehe Kap. 6) für die Session-Verwaltung sowie den Betrieb der Einkaufswagen eingesetzt. Die Bestellungen selber werden im Dokumentspeicher abgelegt (Kap. 6), die Kunden- und Kontoverwaltung erfolgt mit einem relationalen Datenbanksystem.

Bedeutend für den erfolgreichen Betrieb eines Webshops ist das Performance Management. Mit der Hilfe von Web Analytics werden wichtige Kenngrößen (Key Performance Indicators) der Inhalte (Content) wie der Webbesucher in einem Datawarehouse aufbewahrt. Mit spezifischen Werkzeugen (Data Mining, Predictive

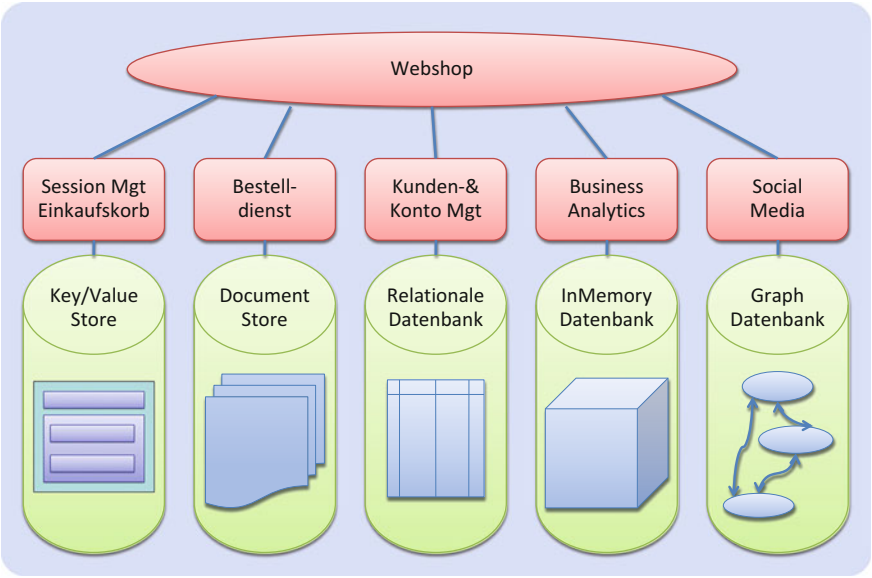


Abb. 2.1 Nutzung von SQL- und NoSQL-Datenbanken im Webshop nach Meier und Kaufmann 2016

Business Analysis) werden die Geschäftsziele und der Erfolg der getroffenen Maßnahmen regelmäßig ausgewertet. Da die Analysearbeiten auf dem mehrdimensionalen Datenwürfel (Datacube) zeitaufwendig sind, wird dieser InMemory gehalten (vgl. Kap. 6 und 10).

Die Verknüpfung des Webshops mit sozialen Medien drängt sich aus unterschiedlichen Gründen auf. Neben der Ankündigung von Produkten und Dienstleistungen kann analysiert werden, ob und wie die Angebote bei den Nutzern ankommen. Bei Schwierigkeiten oder Problemfällen kann mit gezielter Kommunikation und geeigneten Maßnahmen versucht werden, einen möglichen Schaden abzuwenden oder zu begrenzen. Darüber hinaus hilft die Analyse von Weblogs sowie aufschlussreicher Diskussionen in sozialen Netzen, Trends oder Innovationen für das eigene Geschäft zu erkennen. Falls die Beziehungen unterschiedlicher Bedarfsgruppen analysiert werden sollen, drängt sich der Einsatz von Graphdatenbanken auf (vgl. Abschn. 2.3.2 resp. Kap. 6).

2.2 Semantische Modellbildung

Unter Modellbildung versteht man die Abstraktion eines Ausschnitts der realen Welt oder unserer Vorstellung in Form einer formalen Beschreibung. Ein semantisches Datenmodell bezweckt, die Datenarchitektur eines Informationssystems unter Berücksichtigung der semantischen Zusammenhänge der Objekte und Beziehungen zu erfassen. Konkret werden Daten und Datenbeziehungen durch Abstraktion und

Klassenbildung beschrieben. Ein Modellzyklus erlaubt, einmal erkannte Objekte der Realität oder Fantasie und deren Beziehungen untereinander in die reale Umgebung zurückzuführen. Durch die mehrmalige Anwendung der Modellzyklen eröffnen sich Erkenntnisse und Zusammenhänge (kognitive Struktur).

Peter Pin-Shan Chen vom MIT in Boston hat 1976 in den Transactions on Database Systems der ACM sein Forschungspapier ‚The Entity-Relationship Model – Towards a Unified View of Data' publiziert (Chen 1976). Er unterscheidet dabei Entitätsmengen (Menge von wohlunterscheidbaren Objekten der realen Welt oder unserer Vorstellung) und Beziehungsmengen. Entitätsmengen werden als Rechtecke und Beziehungsmengen als Rhomben grafisch dargestellt, wobei die Eigenschaften (Attribute) diesen Konstrukten angehängt werden. Chen stellt gleich zu Beginn seines Forschungspapiers fest: ‚[This model] incorporates some of the important semantic information about the real world …'.

Im Folgenden testen wir das Entitäten-Beziehungsmodell auf die Nützlichkeit sowohl für die Modellierung von SQL- wie NoSQL-Datenbanken. Als Ausschnitt der realen Welt verwenden wir ein kleines Anwendungsbeispiel aus der Filmwelt mit Darstellern und Regisseuren. Ein rudimentäres Informationssystem soll Filme durch Titel, Erscheinungsjahr und Genre charakterisieren. Zudem interessieren wir uns für Schauspieler mit Namen und Geburtsjahr; ebenso für Regisseure. Das Informationssystem soll nicht nur Auskunft geben über Filme und Filmemacher, sondern auch aufzeigen, wer welche Rollen bei Filmprojekten eingenommen hat.

Objekte oder Konzepte der realen Welt werden im Entitäten-Beziehungsmodell als Entitätsmengen dargestellt. Aus diesem Grunde definieren wir in Abb. 2.2 die Entitätsmengen FILM, DARSTELLER und REGISSEUR mit ihren jeweiligen

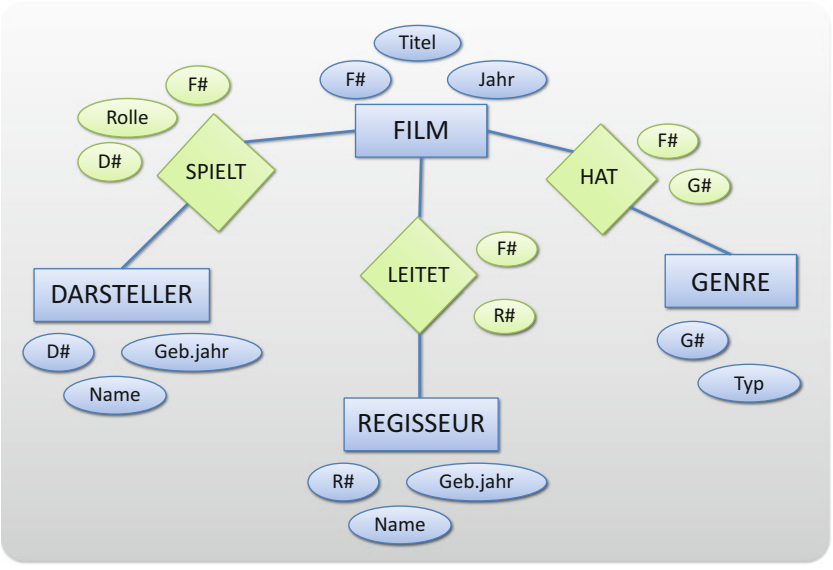


Abb. 2.2 Entitäten-Beziehungsmodell für Filme und Filmemacher

Merkmalen. Da das Genre des Films von Bedeutung ist, entscheiden wir uns für eine eigenständige Entitätsmenge GENRE, um die Filme typisieren zu können.

Beziehungen unter den Objekten werden als Beziehungsmengen modelliert, wobei die Eindeutigkeit einer Beziehung in der Menge durch Schlüsselkombinationen ausgedrückt wird. Die Beziehungsmenge SPIELT besitzt gemäß Abb. 2.2 die beiden Fremdschlüssel F# aus der Entitätsmenge FILM resp. D# aus DARSTELLER. Zudem enthält die Beziehungsmenge SPIELT das eigenständige Beziehungsmerkmal ‚Rolle', um das Auftreten des jeweiligen Schauspielers in seinen Filmen ausdrücken zu können.

Das Entitäten-Beziehungsmodell unterstützt den konzeptionellen Entwurf und besticht, weil es für unterschiedliche Datenbankgenerationen verwendet werden kann. Im ursprünglichen Forschungspapier hat Peter Chen aufgezeigt, wie sich ein Entitäten-Beziehungsmodell in relationale oder netzwerkartige Datenbanken abbilden lässt.

In der Praxis können bei der Entwicklung eines Informationssystems die Bedürfnisse der Anwender im Entitäten-Beziehungsmodell unabhängig von der Datenbanktechnik ausgedrückt werden. Neben der Datenstruktur mit Entitätsmengen und Beziehungsmengen lassen sich Fragen diskutieren, die später durch das Informationssystem beantwortet werden. Auf unser Filmbeispiel bezogen könnte interessieren, in welchen Filmen der Schauspieler Keanu Reeves aufgetreten ist, welche Rollen er jeweils hatte oder ob er bereits als Regisseur Erfahrungen gesammelt hat.

Objekte der realen Welt lassen sich meistens durch Substantive ausdrücken; sie werden durch Entitätsmengen dargestellt und durch Rechtecke symbolisiert. Beziehungen zwischen Objekten werden mit Verben charakterisiert. Chen wählte in seinem Modell für diese Beziehungsmengen ebenfalls ein eigenes Konstrukt (Rhombus). Frage: Was in unserem Leben lässt sich nicht durch Objekte und Objektbeziehungen ausdrücken?

Eine Diskussion des Entitäten-Beziehungsmodells mit den Auftraggebern, mit künftigen Nutzern verschiedener Fachbereiche oder mit Kunden und Lieferanten verifiziert die Datenarchitektur, bevor teure Investitionen in Infrastruktur und Personal geleistet werden. Hinzu kommt, dass die Vertreter unterschiedlicher Anspruchsgruppen sich nicht in den vielfältigen Datenbank- oder Softwaretechnologien auskennen müssen. Ein Entitäten-Beziehungsmodell verwendet natürlich-sprachliche Begriffe (Substantive, Verben, Eigenschaften) bei der Lösungssuche und benötigt keine Übersetzung des Untersuchungsgegenstandes (Universe of Discourse).

2.2.1 Relationenmodell

Das Relationenmodell wurde vom englischen Mathematiker Edgar Frank Codd konzipiert und 1970 unter dem Titel ‚A Relational Model of Data for Large Shared Data Banks' bei den Communications of the ACM veröffentlicht (Codd 1970). Er arbeitete zu dieser Zeit am IBM Forschungslabor San Jose in Kalifornien, wo eines der ersten relationalen Datenbanksysteme unter dem Namen ‚System R'